International Journal of Exploring Emerging Trends in Engineering (IJEETE)
Vol. 03, Issue 02, Mar-Apr, 2016 Pg. 47-52 WWW.IJEETE.COM

# SURVEY ON SENTIMENT MINING TECHNIQUES FOR MACHINE LEARNING

[1]Varahasmay. R, [2]Michael Raj. T.F, [3]Meganathan.S
[1, 2, 3] Asst. Prof. in Dept. of Computer Science and Engineering
[1, 2, 3] SRC, SASTRAUNIVERSITY, Kumbakonam

**Abstract: -** Nowadays people share their opinions over the Internet. The growth of the social websites increases the people's opinion towards the social and non-social entities and attributes. In this technological world mining these opinions and applying the sentiment analysis is a challenging task. The implementation of sentiment mining algorithms over the opinions is very much essential to things get classified and to provide a knowledge base for the information retrieval. This objective of this article is to review the sentiment mining approaches and suggests some recommendations to improve the sentiment analysis process for different data sources.

**Keywords:** Sentiment mining, Levels of sentiment mining, Data sources, Machine learning techniques

## I. Introduction

Sentiment mining otherwise known as opinion mining is the process of extracting emotions, thoughts, and way of thinking, mind-set, point of view, attitudes, appraisals and feelings of the people's over products, services, goods, topics, issues, organization etc. People's sentiments are in the form of text data, data set, web, social networks, opinion polls, debate, and forums. In sentiment analysis process contains various levels of analysis such as Document level analysis, Sentence level analysis and Phase Level analysis. The sys-tem analyzes user's review and classifies it to either positive or negative. The data source contains more number of positive and negative comments. Opinions are directly ex-pressed by the people or expressed indirectly. System has to deal with those reviews and extract the opinion accurately. Sentiment analysis contains various processes such as subjectivity detection, sentiment prediction, sentiment summarization, text summarization, feature extraction and detecting the fake review.

## II. Sentiment Mining

Sentiment Mining is a process for tracking the feel of public about certain products, services, goods, topics, issues, organization etc. It also makes a machine to accumulate and categorize opinions.

Opinion extraction is a task to find out the polarity of reviews. Polarity represents positive or negative [20] [21]. Figure 1 shows Opinion Mining Model

Opinion: Opinion represents the feelings, judgments or mind-set of the user's.

Opinion Holder: may be an organization or person who expresses their opinion about any object.
Object: is a real world entity about which the opinion expressed.

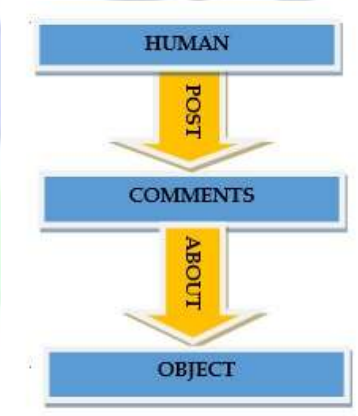


Figure 1 Opinion Model

## III. Levels of Sentiment Mining

### A. Document Level

In Document level analysis, a whole document considered for mining. It classifies that whole document is either positive or negative about any object. It classifies sentiment based on opinion rather than topic. A single review about a topic is considered in document level analysis [14].

Drawbacks:

1. This system does not conclude what person likes or dislikes the object.

  

2. This level of sentiment classification is not suitable for blogs and forums which contain few opinionated sentences.

3. It defines only the polarity of the document. But negative words does not represent the user dislikes every-thing or positive words does not represent the user likes everything.

### B. Sentence Level

Unlike Document level analysis, sentence level analysis considers only sentences contains opinionated terms and states the result that the considered sentences are positive or negative [14].

Example: This is a good camera.

Drawbacks:

1. A user may express more than one feeling in a sentence. If a user express likes as well as dislikes of an object at same sentence leads the system to rank it a neutral. It does not convey the real meaning what the user wants.

2. If user expresses likeness at one sentence and dis-likes at another sentence then this leads to negative result.

### C. Word Level

Word level analysis is otherwise called Phase Level Analysis which considers only the words from a sentence. It tokenizes a sentence into words and extracting the keywords from it then finds whether the keywords are positive or negative.

Example: This is a good phone [14].

Here good is keyword.

## IV. Data Store

Data Source are the location where the data residing. Data Source may be Text File, Review site, Dataset, Social media Comments, Micro-blogging, Blogs, Google play Android Appstore etc.

### A. Text File

A set of positive or negative comments or opinion reviews are stored in the form of text file.

### B. Review Sites

A large number of review sites are available in the internet. It may contains the review of a product, software, restaurant reviews, airway reviews etc. Those form of data used in most of sentiment classification. Some e-commerce websites are www.amazon.com, www.yelp.com, www.CNETdownload.com.

### C. Data set

Dataset available for movie reviews, amazon product re-view, weather forecasting dataset, and tumor cautions re-view dataset are available at websites. Using such a dataset system can perform sentiment polarity classification.

### D. Social Media Comments

Social Media's like Facebook, linked, public forums contains lot of people opinions in the form of electronic text.

### E. Micro Blogging

Short messages sent by the people are represented as Micro-Blog. Twitter messages called "tweets" are used as a data source for sentiment analysis process. Those data source are Micro-Blogging.

### F. Blogs

It is otherwise called Web-Log is a small paragraph of in-formation, opinion, diary called posts arranged in a chronological order. Blogs are used to analyze the mood of public, sales information of movie and sales analysis.

### G. Google Play Android AppStroe

This appstore contain huge amount of reviews about app and its ranking made by the user's opinion/reviews.

## V. Machine Learning Models

Machine learning is the process to make the system to learn by its own. Machine learning algorithms makes system to build a model using some sample data. If any new data arrives then the system can able to predict based on already learned model [3][6][7][16].

### A. Supervised Learning

Supervised learning (machine learning) takes input of known training dataset with labeled classes of the data, and constructs a model that generates the prediction response to new data [3][17].

Bayesian Classifier: Naïve Bayes is classification techniques which follows Bayes theorem that is Bayes statistics. It uses probabilistic approach for classification. Naïve Bayes is otherwise called simple Byes, and Independence Bayes, sometimes it is called as Bayesian classifier and Idiot Bayes.

  

Naive Bayes classifier has been used to classify the text into various categories such as spam or ham, anger, happy, sorrow etc [1][4][8]. Naive Bayes performs well with classes of high dependent feature 3. This approach produces better accuracy when the dimensionality of input is high. For a document d and a class C,

$$Pr(Ri|t) = \frac{Pr(t|Ri)\ Pr(Ri)}{Pr(t)}$$

Where Pr (Ri|t) is the posterior probability

Pr (t|Ri) is the likelihood probability

Pr (Ri) is the prior probability and

Pr (t) is constant value t.

Support Vector Machine: SVM introduced in COLT-92 by Boser, Gyon and Vapnik. SVM can be applied for classification or regression. It follows kernel based approach. SVM classifier has some predetermined format of input and output. The given input can be decomposed into words of vectors. Using those vectors, SVM constructs N-Dimensional hyper plane. The result may either +1 or -1. For each input, SVM predict the class Cj{1,-1} corresponds to positive or negative[3][19].

The document dj as Cj € {1,-1} can be represented weight vector,

$$= \sum \alpha j\ Cj, \qquad \alpha j \geq 0$$

Where, αj is a multiplier.

Maximum Entropy: The Maximum Entropy is a probabilistic classifier which is belongs to the class of exponential models. The maximum entropy does not assume that the features are conditionally independent to each other. It is based on Principle o Maximum Entropy. Maximum entropy follows search based optimization to find weights for the features that maximize the possibility of the training data [3]. The Probability of a class R given a document t and weight λ is

$$P(R|t,\lambda) = \frac{\exp \sum a\ \lambda a\ fa\ (R,t)}{\sum_{R'=R} \exp \sum a\ \lambda a\ fa\ (R',t)}$$

Boosting Algorithm: Boosting is a machine learning collection of meta-algorithm which converts weak learners to strong ones. Boosting involves incrementally constructing an ensemble by training each new instance of model to emphasize the previous model misclassified training instance.

The most popular Boosting algorithm is the Ada-Boost Algorithm. It has been applied to rule-based systems, Bayesian Classifier and decision trees. But one criticism of boosting is, this algorithm perform poorly while classifying the noisy data. Depending on dataset Choosing Boosting algorithm may become unsuccessful [3].

Genetic Algorithms: John Holland invented Genetic Algorithms (GA) during 1960s to 1970s. GA follows heuristic approach which imitates natural selection and survival of the fittest. The solution of Genetic Algorithm is x-bit Chromosome that symbolizes one arrangement. Every chromo-some has a measure of accuracy of the classifier is represented as fitness score. Fitness score for n documents as,

$$Fitness(s) = \sum_{a=1}^{n} \sum_{b=1}^{m} sim(a,a-b)+sim(a,a+b)$$

Where, m is a range that describes the similarity size of neighborhood for each document. Iterations of this algorithm described in following three steps:

1. Take two solutions x and y from the set of all solutions with higher fitness scores.

2. Combine x and y using crossover operator to pro-duce new solution z.

3. Occasionally, mutate solution by exchanging two documents in solution randomly [3]

Bayes Model: Bayes Model is otherwise known as Belief Network or Bayesian Network. Bayes Network is a probabilistic graphical network. It is represented by using directed acyclic graph (DAG). DAG indicates a group of randomly selected attributes and dependencies among them. Each edge represents conditional dependencies; each non connected node represents variables that are not conditionally independent [1][3].

Bayesian Network used by Hernandez and Rodriguez to describe the real world problem by classified it into three related and different variables.

  

They propose Multi-dimensional Bayesian network classifier.

K-Nearest Neighbor Algorithm: KNN is a non-parametric approach used for regression and classification. KNN is a kind of Lazy learners. The input contains K-nearest training examples. Object can be categorized based on maximum number of neighbor's votes. If k=1 then the class can be allocated to the object in case of classification [8] [22].

The example data can be vectors in n-dimensional memory space, each with known class label. KNN memorize the vectors and its known concept or class. K is a constant value defined by the user and it classifies the unlabeled vector by assigning the most frequent label frequent among the k training examples.

### B. Un Supervised Learning

Unsupervised learning is a process of making system to classify data without prior knowledge. It is used to clustering the data into to various groups based on distance or link between them. Unsupervised learning algorithms are used whenever input data set with unknown class labels [3[9].

K-means Clustering Algorithm: MacQueen discovered K-means algorithm at 1967. K-means is famous algorithm used to clustering the unlabeled data. K-means is so easy to cluster the unlabeled data into number of groups (assume k clusters). K-means algorithm group n-attributes into k number of clusters. K-means define n-centroids, each cluster having one centroid. Distinct position of data causes distinct result. Then select each point and group it to the closest centroid. When all the points are taken into the account, then the initial stage is completed and an initial clustering is completed. Now again estimate n new centroids to the newly grouped objects. Then perform the same above mentioned process again and again between the new n-centroids and the input data set. Now, n-centroids move their position little by little until all the clusters are to be done. The objective function is

$$J = \sum_{j=1}^{k} \sum_{i=1}^{n} \| Xi\ (j) - Cj \|2$$

Where, Cj is the centroid of cluster, $\| Xi\ (j) - Cj \|2$ is distance function, J is the objective function, k is the number of cluster and n is the number of cases.

Fuzzy C-Means Algorithm: Dunn invented this algorithm in 1973 and further developed by Bezdek at 1981. Fuzzy c-means is to cluster the unlabeled data together as cluster. In this algorithm same data objects are grouped as many clusters. It uses membership levels to group same data to various clusters. The objective function is described as follows:

$$Jm = \sum_{i=1}^{N} \sum_{j=1}^{C} uijm \| Xi - Cj \|2,\ 1 \leq m < \infty$$

Where, N represent number of data, C represent number of clusters, m is any value which is more than 1, uij is the membership value. cj is the center of the cluster [10][18].

### C. Lexicon Based Approach

Lexicon based approach is based on sentiment lexicon. Sentiment lexicon is a collection of precompiled words and expression related to people mind-set [5][11][18]. Lexicon based approach does not require prior knowledge or prior training to data classification. It is further divided as Corpus based approach; Dictionary based approach, and Manual approach to find sentiment polarity.

Manual Approach: Construction of sentiment and its features are done by manually. It is so tedious as well as time consuming process. It is also impractical task.Dictionary Based Approach: Dictionary based approach initially collect set of opinionated data manually, after in-creasing by searches synonyms and antonyms from Senti-WordNet, Dictionary and thesaurus12. If new words found then it is added to the seed list and do the same process until no new words found. Any correction of errors can be done manually. It does not specify the domain specific opinionated words and its orientation.

Corpus Based Approach: Corpus based approach is based on prototype of corpora or document. To prepare a corpus require more number of words so it is not much effective like dictionary based approach. Corpus based approach does specify the domain specific opinionated words and its orientation.

  

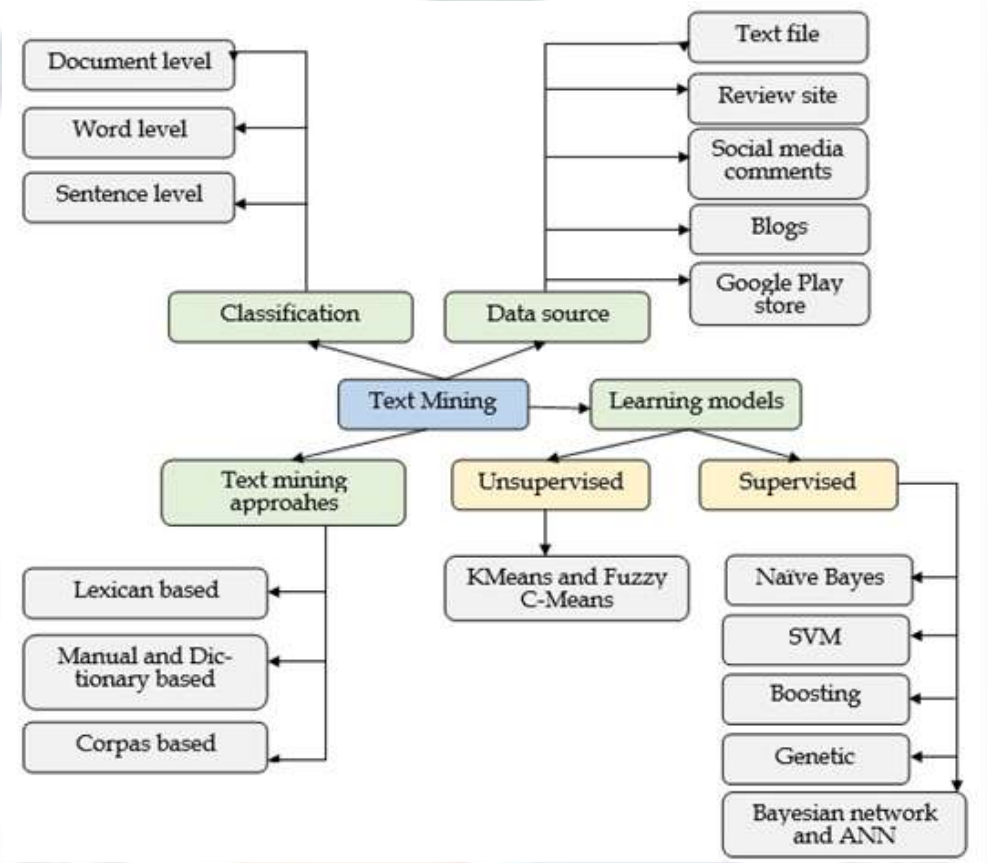


**Figure 2.** Text Mining Levels, Data Source and their Classification

## VI Conclusion and Future Works

The main aim of this survey is to discuss various techniques used for sentiment classification, various levels used in the mining process and various kinds of data sources. It is rep-resented in the above figure 2. Many of the organizations have putting their efforts in finding the best system for sentiment analysis. Some of the algorithms give good results but still many more limitations in these algorithms. This domain requires well scalable algorithm to classify the text accurately. In future it may be extended to mine opinions of all the languages.

## REFERENCES

[1] Anwer N and Rashid A. Feature Based Opinion Mining of Online Free Format Customer Reviews Using Frequency Distribution and Bayesian Statistics. Networked Computing and Advanced Information Management (NCM). Sixth International Conference. 2010 Aug 16: 57-62.

[2] Fan M, WU G. Opinion Summarization of Customer comments. International conference on Applied Physics and Industrial Engineering. 2012; 24:2220-26.

[3] Shelke N M, Deshpande S and Thakre V. Survey of Techniques for Opinion Mining. International Journal of Computer Applications (0975 – 8887). 2012; 57(13):30-35.

[4] Khushboo T. Mining of Sentence Level Opinion Using Supervised Term Weighted Approach of Naïve Bayesian Algorithm. Int. Journal. Computer Technology & applications. IJCTA. 2012; 3(3):987-91.

[5] Wang W, Xu H and Wan W. Implicit feature identification via hybrid association rule mining. ESWA 8310. Model 5G. 2013 July; 40(9):.3518-31.

[6] Ron Kovahi; Foster Provost. Glossary of terms. Machine Learning. 1998; 30: 271–74.

[7] Bishop C M. Pattern Recognition and Machine Learning. Springer, ISBN 0-387-31073-8. 2007 Jan; 17(5): 1-3.

[8] Leo Breiman Bias, Varience, and Arcing classifiers technical report. Retrieved 19 January 2015.

  

[9] Ortigosa-Hernandez Jonathan, Rodriguez Juan Diego, Alzate Leandro, Lucania Manuel, InzaInaki, Lozano, Jose A. Approaching sentiment analysis by using semi-supervised learning of multi-dimensional classifiers. Neuro computing. 2012 sep 1.
[10] Bezdek J C Pattern Recognition with Fuzzy Objective Function Algorithms. Plenum Press. New York.1981.
[11] Kechaou Z; Ben Ammar M; Alimi A M, Improving e-learning with sentiment analysis of users' opinions. Global Engineering Education Conference (EDUCON). IEEE. 2011 April 6: 1032-38.
[12] Miller G, Beckwith R, Fellbaum C, Gross D, Miller K. WordNet: an on-line lexical database. Oxford Univ. Press. 1990; 3(4): 235-44.
[13] Mohammad S, dunne C, Dorr B. Generating high-coverage semantic orientation lexicons from overly marked words and a thesaurus. In: Proceedings of the conference on Empirical Methods in Natural Language Processing (EMNLP'09). 2009; 2:599-608.
[14] Neha S. Joshi Suhasini A. Itkat. A Survey on Feature Level Sentiment Analysis.2014;5(4):5422-25.
[15] Culibrk, Mirkovic M, Lugonja P, Crnojevic V. Mining Web Videos for Video Quality Assessment. Soft Computing and Pattern Recognition (SoCPaR). International Conference. 2010 Dec 7; 75-80.
[16] Ayesha Rashid1, Naveed Anwer2, Dr. Muddaser Iqbal3, Dr. Muhammad Sher4. A Survey Paper: Areas, Techniques and Challenges of Opinion Mining. 2013 Nov; 10(6)2:18-31.
[17] Richa Sharma1, Shweta Nigam2 and Rekha Jain3. Supervised Opinion Mining Techniques: A Survey. 2013; 3(8):737-42.
[18] Sheeba1 J I, Dr. Vivekanandan K. A Fuzzy Logic based on sentiment classification.2014 July; 4(4):27-44.
[19] Abd Samad Hasan Basaria, Burairah Hussina, I. Gede Pramudya Anantaa Junta Zeniarjab. Opinion Mining of Movie Review using Hybrid Method of Support Vector Machine and Particle Swarm Optimization. Procedia Engineering 2013.
[20] Bing Liu. Sentiment Analysis and Opinion Mining. Edition 2. Morgan & Claypool Publishers. 2012 May 167.
[21] Liu B. Web Data Mining: Exploring hyperlinks, contents, and usage data. Opinion Mining. Springer. 2007.
[22] Michael Raj. T.F, SivaPragasam. P, BalaKrishnan. R, LalithambaI. G, Ragasubha. S, QoS Based Classification Using K-Nearest Neighbor Algorithm for Effective Web Service Selection, IECCT, pp.1-4.

  